Assignment 4 - Comparative Training Methods

1. Introduction

Objectives: The purpose of this assignment was to compare three different training methods used in generative AI models:

1. **Pre-training** (Unsupervised Learning)
2. **Supervised Fine-Tuning** (SFT)
3. **Reinforcement Learning** (RL-lite using REINFORCE)

These methods represent the primary stages of training modern large language models like ChatGPT. Pre-training teaches the model general language patterns. Supervised Fine-Tuning teaches the model how to follow instructions. Reinforcement Learning enhances the model by providing feedback and rewards.

Why these methods and settings were chosen

These methods were chosen because they follow the standard pipeline used in modern Generative AI systems: **pre-training, SFT, and RL**. A small character-level GPT model was used because it can be trained quickly on Google Colab while maintaining the learning behavior of larger models.

Hardware and time constraints

The experiments were performed on Google Colab using a GPU. Because of limited computing power and time, a small GPT model was used with:

- 2 layers
- Hidden size 128
- Context length 64

This allowed training to be completed in minutes instead of hours.

2. Methods

Dataset

The **WikiText-2** dataset was used for pre-training.

Dataset size:

- 15,000-25,000 characters (first 2,000 lines used).
- Character-level tokenization.

Preprocessing steps:

- Converted text into character tokens.
- Added special tokens.
- Split into 90/10 for pre-training, 70/30 for SFT.

For Supervised Fine-Tuning, custom instruction-answer pairs were created, such as:

Instruction: Give a creative tagline for coffee

Answer: Sip ideas. Brew brilliance.

Model Architecture

A **Tiny GPT** model was used.

Architecture details:

- Layers: 2
- Hidden size: 128
- Attention heads: 4
- Context length: 64
- Parameters: 525,268

This is a more basic transformer decoder.

Training Settings

Method	Steps	Batch Size	Learning Rate	Optimizer
Pre-training	1,000	32	0.003	AdamW
SFT	200	32	0.0006	AdamW + weight decay
RL	200	N/A (single examples)	0.00015	AdamW

Supervised Fine-Tuning (SFT) Instruction Construction

Instruction-answer pairs were created manually.

Examples:

Instruction: "Give a creative tagline for coffee:"
Answer: " Sip ideas. Brew brilliance."

Instruction: “Write a short motto about learning”
Answer: “Learn, iterate, and grow.”
These examples teach the model how to respond properly.

RL Reward Function

A custom reward function was designed based on:

- **Length**: +1.0 if 10-120 characters
- **Keywords**: +0.3 per relevant word ("learn", "neural", etc.)
- **Diversity**: +0.5 for character variety
- **Penalty**: -0.1 for double spaces

Example: Reward increased if response included words like **model, learn, data.** Reward was reduced for repetitive or poor text.

3. Results

Pre-training Results

Results:

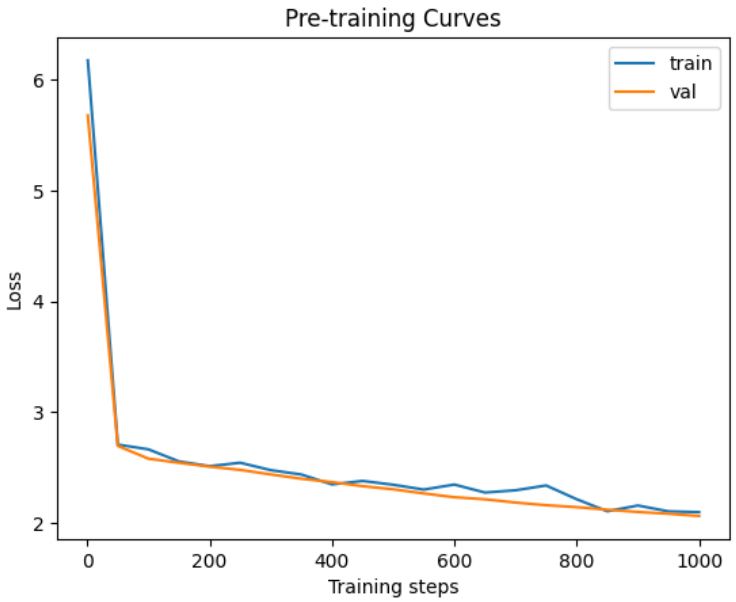
- **Train Loss:** 2.0983
- **Validation Loss:** 2.0638
- **Perplexity:** 7.88
- **BLEU Score:** 0.0469

Observations:

- Smooth convergence.
- Train and validation losses stay close (2.1).
- No overfitting.

This means the model successfully learned general language patterns.

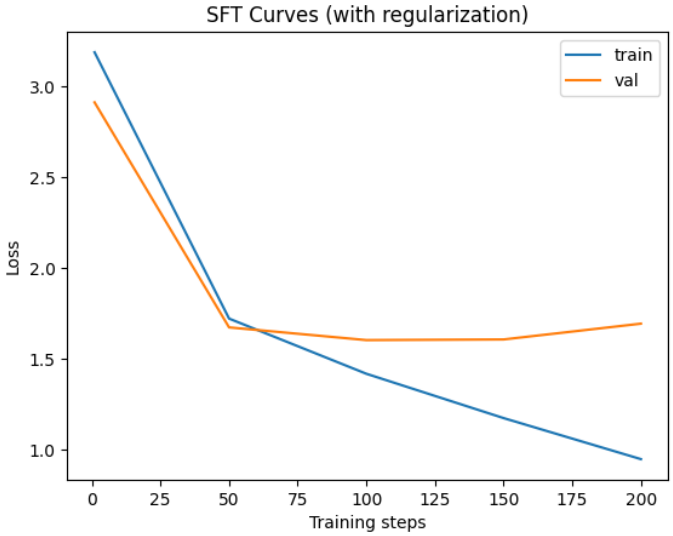
Loss curve:



Supervised Fine-Tuning (SFT) Results

Results:

- **Train Loss:** 1.0086
- **Validation Loss:** 1.7101
- **Perplexity:** 5.53
- **BLEU Score:** 0.045



Observations:

- Training loss decreased significantly.
- Validation loss decreased but remained higher than training loss.
- This shows some overfitting.
- However, perplexity improved compared to pre-training.

RL Results

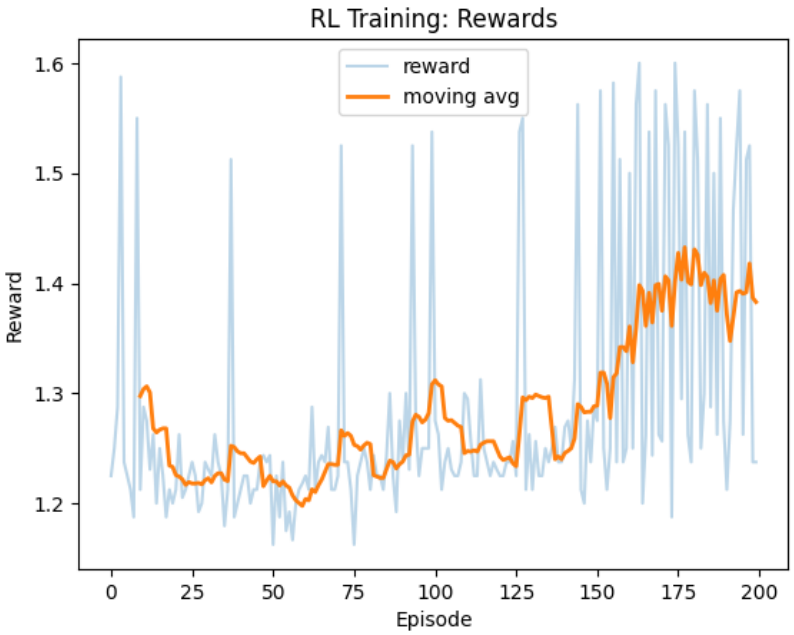
Results:

- **Average reward:** 1.46
- **Baseline:** 1.50
- **BLEU Score:** 0.0670

Observations:

- Reward increased slightly.
- However, improvement was small.

Reward graph:



Perplexity & Bleu Comparison Table

Method	Val Loss	Perplexity	Bleu
Pre-training	2.0638	7.88	0.0469
SFT	1.7101	5.53	0.045
RL	N/A	N/A	0.0670

Qualitative Results

```
Prompt: Instruction: Give a creative tagline for coffee:
Answer:
--------------------------------------------------
Pre-train: Instruction: Give a creative tagline for coffee:
Answer:rot rans , hic on be is thad ssen orties he at comm. The for
SFT:       Instruction: Give a creative tagline for coffee:
Answer:ó. Give)
Instruction: Answer: Study trouction:
Answer: Sut i
RL:        Instruction: Give a creative tagline for coffee:
Answer:1 Wersterster: Whatel model memodel model model model playl

Prompt: Instruction: Write a short motto about learning:
Answer:
--------------------------------------------------
Pre-train: Instruction: Write a short motto about learning:
Answer: panalaly proy deary ) on an 196832 : a secupt preetions ov
SFT:       Instruction: Write a short motto about learning:
Answer:s Suct)
Instruction: easuction: Explain: Sip tudy itive a mo
RL:        Instruction: Write a short motto about learning:
Answer:1 Wh Ster: 'Atttencenssmenswer: Whattel nemerel beearn melog

Prompt: Instruction: Explain what is perplexity:
Answer:
--------------------------------------------------
Pre-train: Instruction: Explain what is perplexity:
Answer:á , ( , araren he . Unertical . Rec . O ) . Lestitional . AS
SFT:       Instruction: Explain what is perplexity:
Answer:r Wrion: iver:
Answer: Exity a creaswer: Warthat play (1 sen
RL:        Instruction: Explain what is perplexity:
Answer:1 ExWh WhelGal Thel model model bearn your model modadl nemo
```

4. Discussion

Training Stability

1. **Pre-training:** Very stable, Smooth convergence
2. **SFT:** Fast convergence Some overfitting
3. **RL:** Unstable, Reward fluctuations

Transfer Effects

- Pre-training provided basic language understanding.
- SFT successfully improved instruction-following ability.
- This shows transfer learning works effectively.

RL Strengths and Weaknesses

Strength: Reward increased slightly.

Weakness: Model exploited reward function.

Instead of generating better text, it repeated keywords. This is called **reward hacking**.

Resource and Time Cost

Training was fast because the model was small.

- **Pre-training:** 8-9 minutes
- **SFT:** 2 minutes
- **RL:** 2 minutes

If scaled up: A larger dataset, more training steps, and a better reward function would all help improve performance.

5. Conclusion

We successfully implemented a three-stage training pipeline that included pre-training, supervised fine-tuning, and reinforcement learning. Pre-training helped the model to learn general language patterns. SFT improved the model's ability to follow instructions while reducing perplexity. Reinforcement learning increased rewards slightly, but it had limitations due to reward design issues. These results show that training and aligning language models is complex. Modern AI systems require large datasets, better reward functions, and significant computing power. **Supervised fine-tuning** achieved the best balance of performance and training stability.